Data Preservation Overview and Discussion

Jan Lundy
jflundy@raytheon.com
(520) 545-8062

Denise Duncan
dduncan@lmi.org
(703) 917-7378

Agenda

- What is data preservation?
- Why do we need to concern ourselves with this?
- Who needs to be involved?
- When should we start thinking about it?
- What’s the state of the practice? What’s ‘bleeding edge’?

What is data preservation?

- Methods and tools to keep authentic data products available for future use
- Authentic
  - Identical in essential respects
    - ‘Essential respects’ defined by future use requirements
  - Issues of provenance
- Data products
  - Reports
  - Views
  - Models and the data to run them
  - Financial data and reports
  - Etc.
- Available
  - Accessible
  - Able to recreate in a usable form

Motivations for data preservation –why?

- Short- to mid- term: Keeping relevant/legally required data on hand
  - Client, statutory, or regulatory requirements
  - Convenience, practicality, smart business process support
- Mid- to long-term: Retention of critical business data for:
  - Risk mitigation or management
  - Legal reasons
  - Competitive purposes
  - Customer and product support, including depot
  - KM and data mining

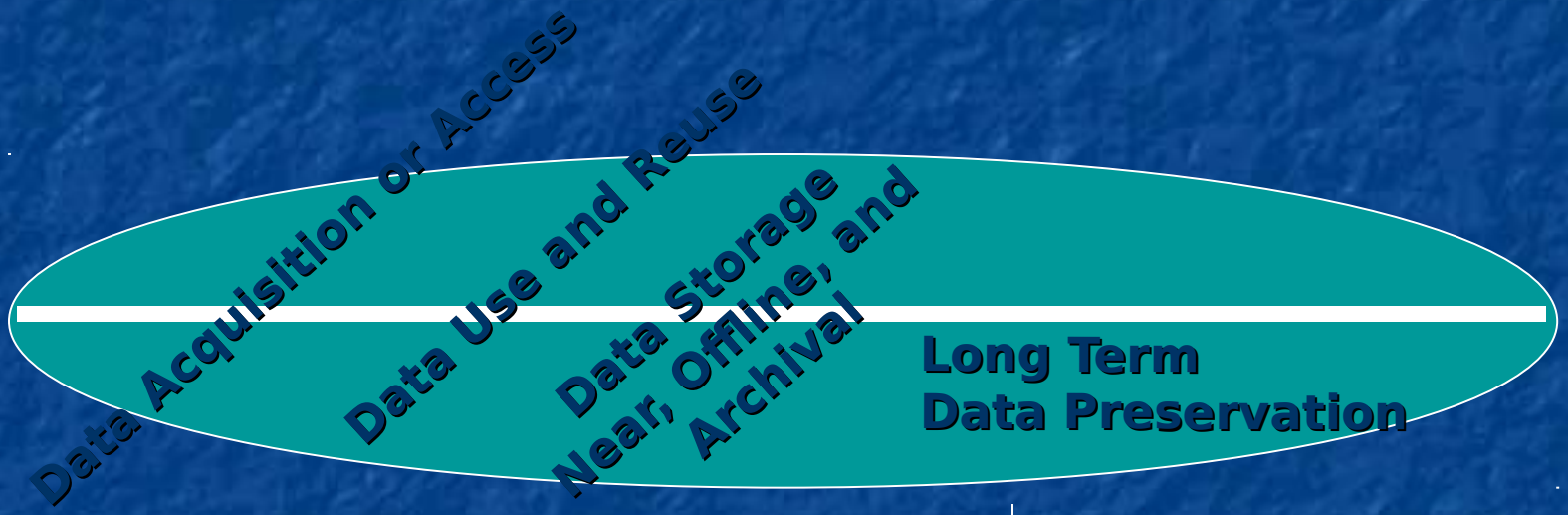
Who should be on the team?

- Awareness and support:
  - Senior management and corporate counsel
- Setting requirements:
  - Customer, Legal department, data users
  - Potential future users
- Performing the job
  - Records Management, DM, Archivist, Librarian and all users

When should you implement data preservation?

- Actually, before any data is generated
  - At requirements stage, develop plan for long term access, use and retention requirements
- Plan for full life-cycle of data
  - Some types of data may be immediately archived; others later in the life cycle
- Planning may alter DM – Records Management relationship

The Data Preservation Timeline



Data Strategy
Concept of Operations
Risk Assessment & Management
Records Management

Media
Read devices
Migration steps

Current Practices

- Technology preservation
- Paper storage
- Migration
- Emulation

Technology Preservation

- Literally, maintain original technology that reads/manipulates the data
  - Hardware
  - Operating system
  - Peripherals
  - Application software

Paper Storage

- Well-known method, readily implementable
- Easily contracted out to third-party providers
- Works well for objects that are useful in non-digital (analog) form (documents, images)
  - Usability compromise--some objects must be in digital form (models, interactive objects)

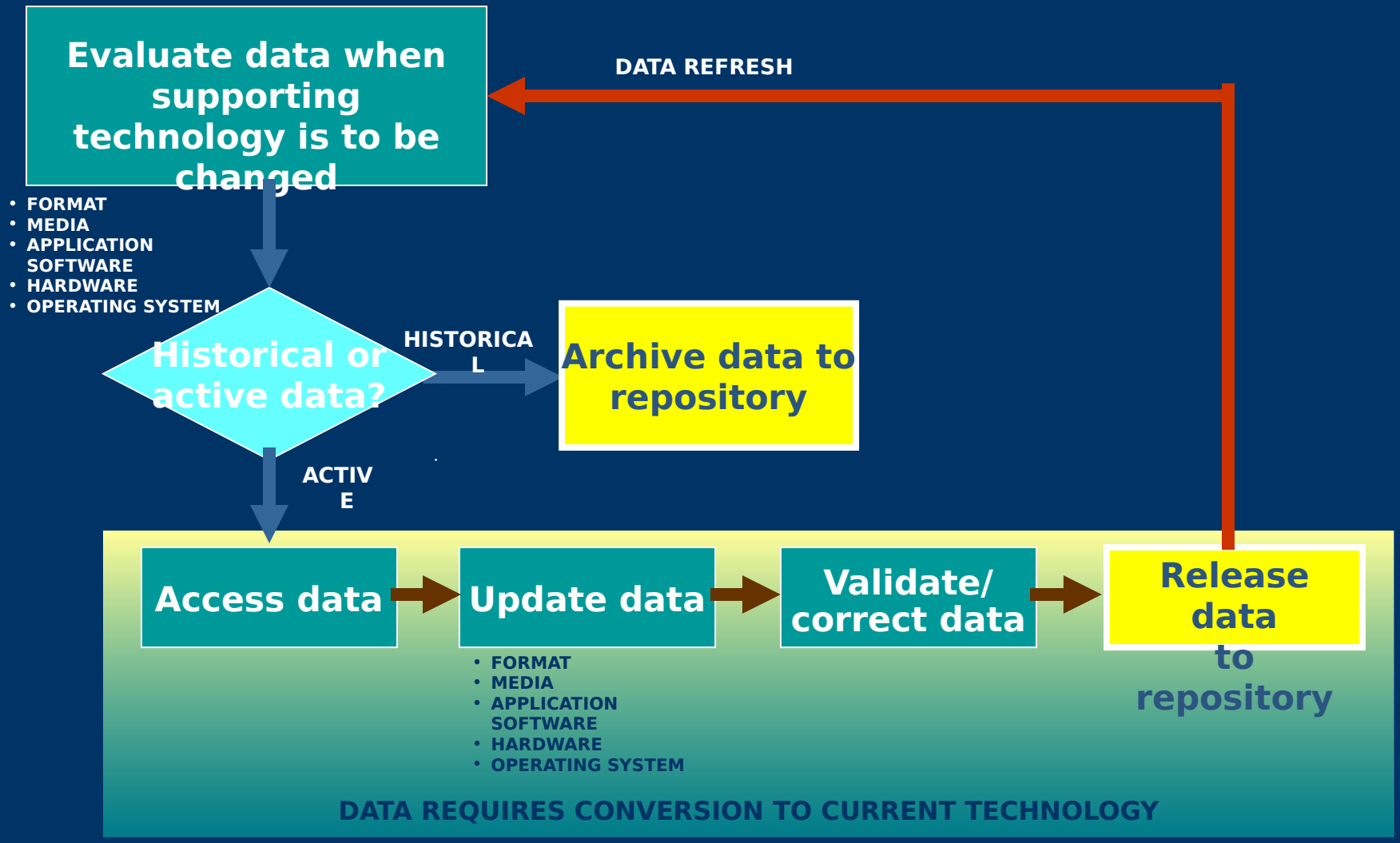
Migration

- Must maintain a current version of complete environment and synchronize:
  - Operating system
  - Application(s)
  - Data (and metadata)
  - Hardware
  - Storage media
- One simplifying variation: migration to a standard format
  - PDF, XML, TIFF, etc.

Emulation

- Keep data in original format
- Maintain ‘equivalent’ technology to manipulate the preserved data
  - Operating environment = state of the art technology
  - Data = original
- Over time, this will turn into migration approach for the operating environment

Data Archive/Preservation Process



New and Emerging Approaches

- Permanent formats (PDF-A)
- OAIS (proposed ISO standard for archives)
- Universal Virtual Machine
- Tools for migration:
  - Typed Object Model
  - Object Interchange Format
  - Rosetta Stones
- Persistent Archives
  - NARA

Parts of the Preservation Solution

- Preservation Planning
- Collecting data objects for archiving
- Maintaining the archive
- Deciding when preservation no longer needed

Cost Models

- Data preservation is still in the ‘one-off’ stage of development
- Cost drivers depend on requirements:
  - For legacy data, selection of items, addition of metadata may be cost drivers
  - Data formats can cause complexity in solutions, resulting in no economy of scale
- Some good studies started:
  - Online Computer Library Center (OCLC): http://www.dpconline.org/graphics/events/presentations/pdf/BellingerDPCForum_CostsBusinessModels.pdf
  - British Library: http://www.dpconline.org/graphics/events/presentations/pdf/LifecycleDPC.pdf

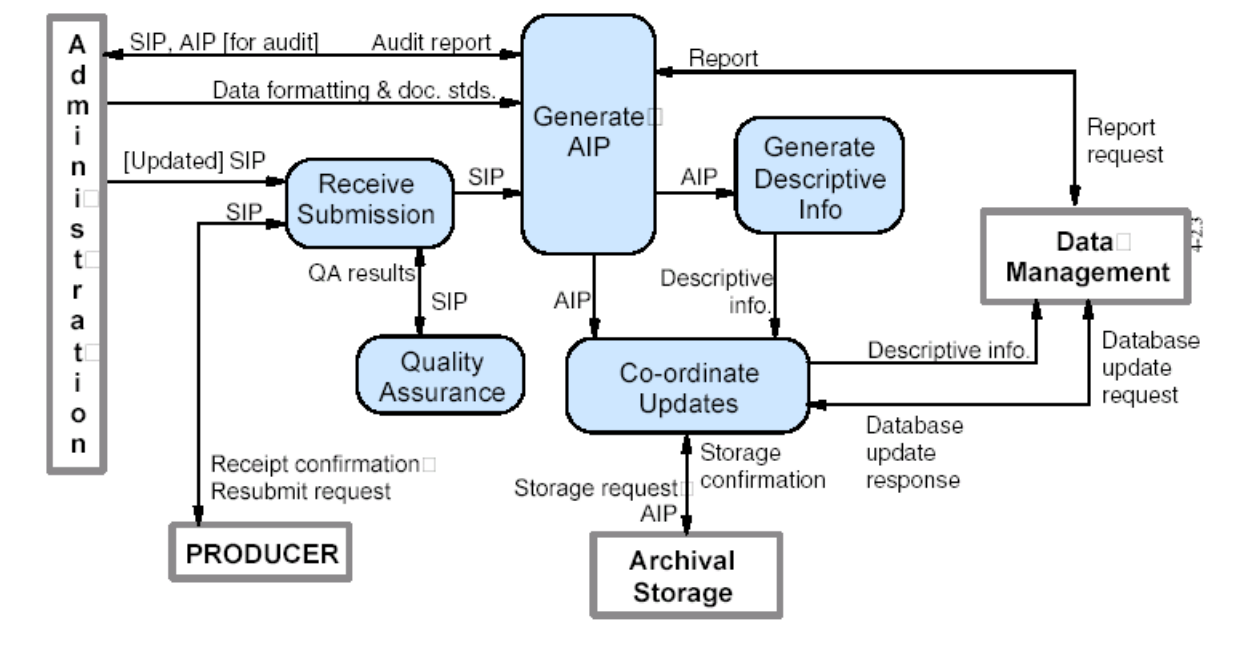
Data Preservation Resources

- International Research on Permanent Authentic Records in Electronic Systems
  - http://www.interpares.org
- NARA's Electronic Records Archives project
  - http://www.archives.gov/electronic_records_archives/about_era.html
- Preserving Access to Digital Information
  - PADI (http://www.nla.gov.au/padi/index.html)

Reference Model for an Open Archival Information System (OAIS)

- A framework for archival systems to preserve and maintain access to digital information over the long term
- Genesis in the library and space science communities
- Useful for the ‘big picture’, but not a step-by-step implementation guide
- The official web site for OAIS activities is http://ssdoo.gsfc.nasa.gov/nost/isoas/

Example—the OAIS *Ingest* process



The Big Picture

- Preservation systems not a commodity item yet
  - Each approach is customized to requirements
  - Technological change spawns new approaches constantly
- Best approach is to scan the literature, understand the current issues and plan according to requirements and current state of the art.